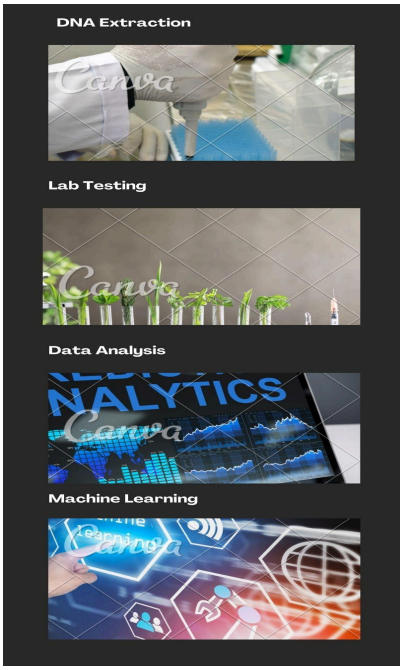
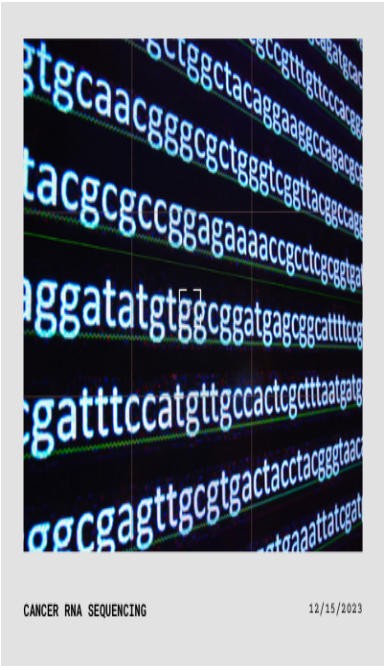
Independent Project Report on

Identification of prognostic biomarkers for prediction of survival risk in cancer patients using lncRNA expression.





Reegina Tyagi

ABSTRACT

This study aims to identify prognostic biomarkers and develop the models for the prediction of the risk associated with the survival in head and neck squamous cell carcinomas (HNSCC) patients using lncRNA. This is accomplished by determining the role of lncRNA expression of 3728 genes in 500 patients in the context of patients' survival time and survival status.

The methods used in the study begins from data collection and preprocessing . RNAseq data for head and neck squamous cell carcinoma (HNSCC) is obtained from 'The Cancer Genome Atlas' (TCGA) . Data was filtered, censored and normalized using the package "DESeq2 in R before performing survival analysis . We obtained the hazard ratios (HR), concordance index(CI) and p values of expression data through coxph function from package "survival" in R . Prognostic indexes (PI) were determined and a voting model was created separating the high risk groups and low risk groups by setting N/2 as cut off (N is the number of total genes). Genes with p values < 0.05 for expression data were sorted from high HR values to low and were divided into two groups the top five were called BPM (bad prognostic markers) as their HR values were high and bottom five with lowest hazard ratio values were called GPM (good prognostic marker). Now PI of these BPM and GPM genes were determined and a voting model was created. The '*sklearn*' package in Python was used to construct the regression models that were implemented to fit the expression values of the BPM and GPM genes against the OS time. Regressors including Linear, K-nearest neighbours, Elastic Net, Random forest, Lasso Lars, Ridge, and Lasso were used along with LazyRegressor. A fivefold cross-validation scheme was carried out for the fitting and evaluation of the test.

INTRODUCTION

The beginning of most HNSCC is in the mucus membranes that line the inside of the mouth, nostrils and throat. Squamous cells make up these membranes so the cancers growing within these cells are referred to as squamous cell carcinomas. Head and neck squamous cell carcinoma is the 6th among the category of most common cancers in the world with 450,000 reported deaths and 890,000 rising cases in 2018. Mostly affecting people over the age of 50 and in men the rate is twice as women. For diagnosis, 66 years is the median age of non-viral HNSCC, while it is 53 years and 50 years respectively for human papillomavirus (HPV)-positive and Epstein–Barr virus (EBV)-related cancers.

The excessive occurrence of these in areas along with Australia and Southeast Asia is related to intake of carcinogenic products like betel quid, aforementioned tobacco, areca nut, alcohol, smokeless tobacco and slaked lime . At the molecular level, HNSCC related to HPV is different from HNSCC associated with tobacco (HPV-negative) and their treatment should be according to their subtype.

HPV harbors oncoproteins E6 and E7 (causes HNSCC) which are linked to increase in copy number of somatic variants. TRAF3 gene is either shortened or deleted in HPV- positive HNSCC and it codes for a protein that is responsible for the regulation of immune response. Production E2F transcription factor family members involved in cell cycle regulation are increased with the amplification of E2F1 gene. Mutations in PIK3CA can also be seen.

In tumors that are not associated with HPV amplifications of 11q13 and 11q22 can be seen that lead to an interaction between BIRC2 and FADD, due to which cell death is inhibited. In tumors which are associated with smoking there are TP53 mutations, inactivation of CDKN2A, and alterations in copy number.

A prognostic marker is analyzed before treatment starts and this identifies tumor specific molecules or histopathological characteristics which includes germline or somatic mutations , changes in microRNA levels ,DNA methylation or circulating tumor cells in blood level that are associated with long term outcome . Prognostic biomarkers help for the selection of patients that require more intensive surveillance or adjuvant therapy. Cytogenetic abnormalities can be taken as prognostic biomarkers for diagnosis of acute myeloid leukemia(AML) . Inversions , deletion and translocation can lead to favorable prognosis like inversions in chromosome 16 , translocation between chromosome 8 and 21 as well as translocation in chromosome 15 and 17, these all are associated with an unfavorable prognosis. By the help of prognostic markers we can characterize various cancers and their associated survival risk . Unusual tissue or cell structure could also be used as prognostic marker.Tumour nodal status, size, presence or absence of lymphovascular invasion and size are commonly used as prognostic markers in breast cancer.

METHOD

Dataset and preprocessing

RNAseq data for HNSCC was obtained from The Cancer Genome Atlas' (TCGA) using the R package "*TCGAbiolinks*". For filtration of data the colData() and rowData() functions from package "*SummarizedExperiment*" were used. *colData()* extract the sample metadata , here it is the information of patients from the array of datasets. After getting the data of patients we removed all the normals and kept the details of patients with all the stages of tumor. For censoring the column containing information of days_to_death was merged with column containing information of days_to_last_follow_up and the patients with NA values in the merged column were removed. Now we had data of 500 patients. The vital status was turned to "0" for "Dead", "1" for "Alive". rowData() extracts gene data containing Ensembl IDs and gene names where we filtered out the data for 3728 lncRNA coding genes using the R package "*biomaRt*". assay() function was used to extract the expression data of 3728 genes for 500 patients that we got after filtration and censoring. The expression data was normalized using "DESeq2".

Survival analysis

Cox proportional hazards (Cox-PH) regression was used for the screening of survival-related genes from their lncRNA expression data with the help of formula:

$$h(G,t) = h_0 \times e^{\beta G}$$

Here the variable G is the gene expression value , h is the hazard function and the variable t is the overall-survival time. The implementation of Cox-PH regression is done with the help of R packages '*survival*' and '*survminer*'.

The Cox regression coefficient, $\beta > 0$ means, the high gene expression is not favorable for the survival and for $\beta < 0$ inverse applies. The hazard ratio (HR) ,$HR > 1$ means if the gene expression is increased there is an increased risk of death for the patient, and for $HR < 1$ the inverse applies. However, if $HR = 1$ there is no risk or impact due to the expression. Survival related genes have been recognized with HR more than or less than 1 and p value < 0.05. The metric Concordance index (C) was used to evaluate the predictive performance of the model.

Prognostic Index

The PI for a different set of genes was used for segregating risk groups, and standard metrics such as HR, *p* value, etc. were found to estimate model's performance by using univariate Cox-PH regression model. The formula used to evaluate PI for a set of k genes is given as:

$$PI_K = G_K \beta_K$$

Here β is the regression coefficient obtained for a gene k, and G here carries the gene expression values for the k genes. The cutp function from '*survMisc*' package in R was used to find the cut-off value for PI values. Patients with a PI value greater than the cut-off were said to be at high risk and were labeled so, whereas patients with a PI value less than the cut-off were said to be at low risk and labeled accordingly.

Voting model

For an n-gene voting model, each patient sample is assigned a vector of length n. Corresponding to the median cutoff of gene expression values, 'high' and 'low' labels are given for values higher than cut-off and lower than cut-off respectively and these labels are counted and the samples are allotted 'high risk' or 'low risk' as an overall label based on the dominant 'label'. Dominant 'label' means the 'label' that occurred for more than $n/2$ times. Thus, each patient was denoted by a vector of n number of risk labels for n number of survival-associated genes.

Machine Learning-Based Regression Models

The '*sklearn*' package in Python was used to construct the regression models. Regressors including Linear, K-nearest neighbours, Elastic Net, Random forest, Lasso Lars, Ridge, and Lasso were used. The models were implemented to fit the expression values of the 10 genes against the OS time. The expression values were independent variables against the 'survival time' was the dependent variables. A fivefold cross-validation scheme was carried out for the fitting and evaluation of the test.

RESULTS

The HR values that were found for gene expression of 3728 genes were sorted from high to low and genes that had p values < 0.05 for their expression values were filtered and top five and bottom five genes were selected from those filtered genes as BPM (genes with HR> 1) and GPM (gene with HR< 1) respectively for the study. The dataset with Hazard Ratio, P value for those values is as follows:

Gene	Hazard_Ratio	p_value	Coefficient
LINC01810	4.136856566	0.046049651	1.419936216
LINC02113	3.287686317	0.000138253	1.19018407
XIRP2-AS1	2.401773079	0.010298402	0.876207248
FAM197Y6	2.346977498	0.018442406	0.853128329
LINC01650	2.330760409	0.006470601	0.84619457
LINC01218	2.182492438	0.004883727	0.780467544
LINC01742	2.142106247	0.004801893	0.761789572
LINC02529	1.98457941	0.001090701	0.685407007
LINC01626	1.865433827	0.029828317	0.623493641
MYO16-AS2	1.835760756	0.007337384	0.607458977

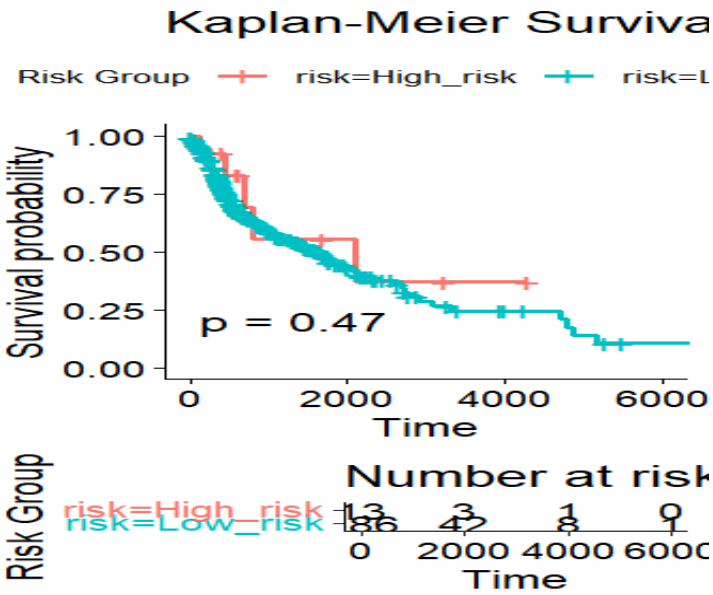
Here above five are BPM and bottom five are GPM. Coefficient is cox regression coefficient, high expression count is number of gene expression values above median for 500 patients , low expression count is for number of gene expression values below median for 500 patients.

The expression values of BPM in 500 patients were used to find HR value, concordance index , p value , high expression count , low expression count.

Gene	Hazard_Ratio	p_value	Coefficient
LINC00424	0.32432193	0.01302408	-1.126018645
GPC5-IT1	0.375170032	0.017944734	-0.980375936
LINC00596	0.379819288	0.007301275	-0.968059698
LINC01685	0.435591004	0.011104152	-0.83105154
LINC02368	0.479468129	0.017840006	-0.735077854
LINC01627	0.488014991	0.035509922	-0.717409154
LINC01570	0.528605007	0.017651384	-0.637513804
LINC01307	0.528750449	0.012424394	-0.6372387

RPS6KA2-AS1	0.530091504	0.004881606	-0.634705638
LINC02325	0.534363977	6.59E-06	-0.626678068

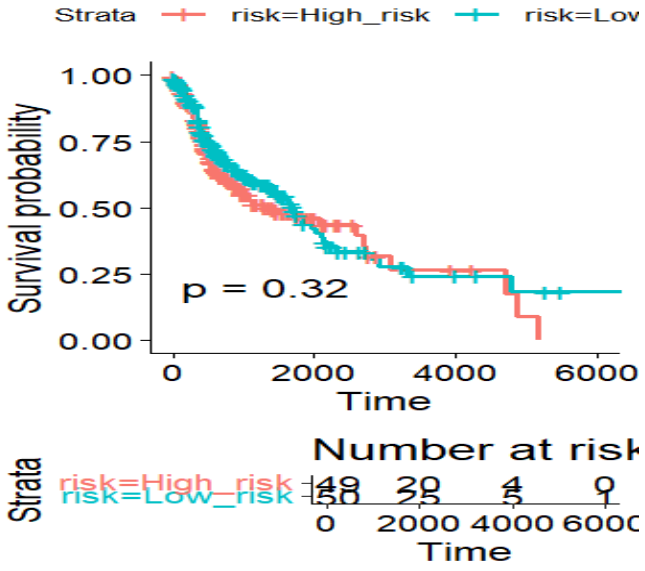
Kaplan-Meier plot for voting model:



Results of PI Model:

	Hazard_Ratio	p_value	Concordance_Index	Log-rank test p-value
PI_model	1.66903537	0.000235103	0.5741790387	0.3247194961

Kaplan-Meier plot for PI model:

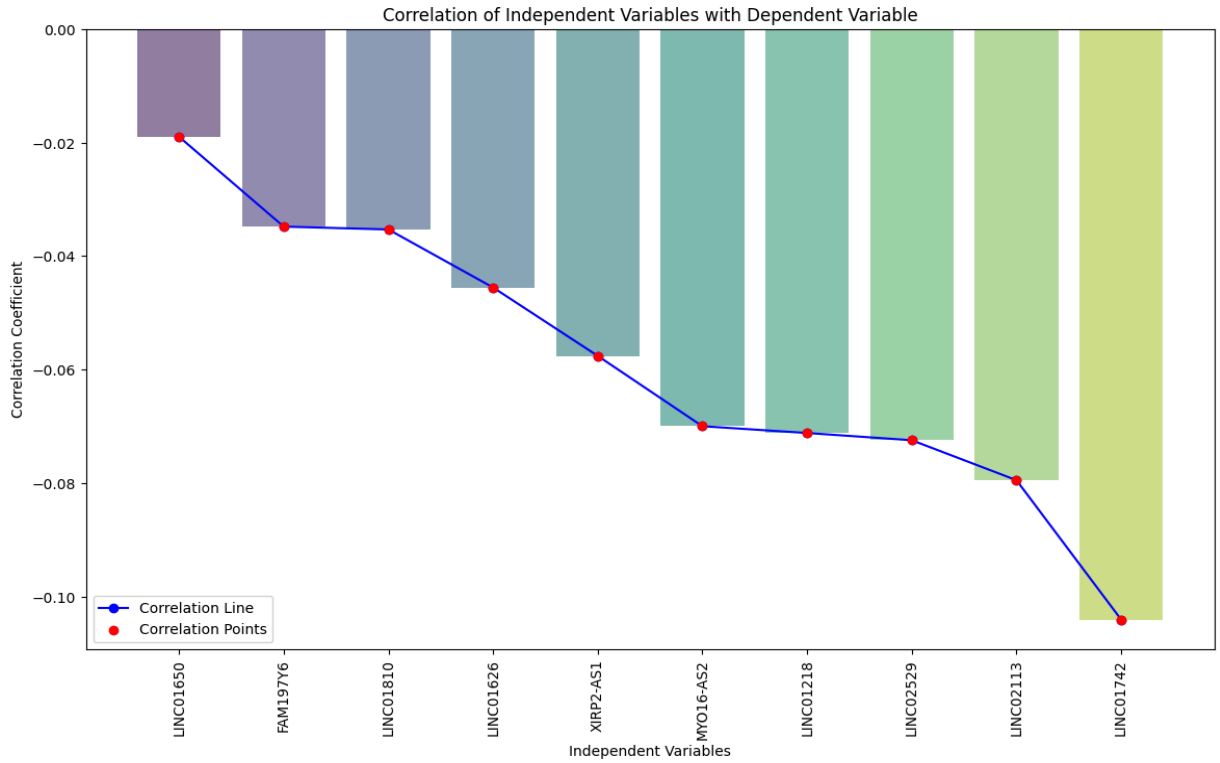


Regression methods were implemented to fit the expression values of the BPM and GPM genes against the OS time (days_to_death) that was converted into years. The data of these genes was split initially into 80% training data and 20% testing data separately. The results we got using the training data for five fold cross validated models were the metric scores we found as mean of mean absolute error, root mean score error, r2 score.

For BPM:

Regressor	MSE	R2
LinearRegression	6.116344677	0.03203529006
RandomForestRegressor	7.207829008	0.04079933357
Ridge	7.357061755	0.04244143118
Ridge {'alpha': 10, 'fit_intercept': True, 'max_iter': 1000, 'tol': 0.0001}	7.355653639	0.0426247045
PoissonRegressor	7.3441	0.05

The gene expression showed a negative correlation with OS, Here is the curve:



For GPM:

Regressor	MSE	R2
LinearRegression	6.021123025	0.06206470139
Ridge	6.011443583	0.06357250814
Ridge {'alpha': 10, 'fit_intercept': True, 'max_iter': 1000, 'tol': 0.0001}	5.95253485	0.07274896576
BayesianRidge	5.9536	0.08

GammaRegressor	5.9049	0.08
LassoLarsIC	5.9049	0.08
MLPRegressor	5.8564	0.09
OrthogonalMatchingPursuitCV	5.8564	0.09
PoissonRegressor	5.8081	0.1
OrthogonalMatchingPursuit	5.6644	0.12

The gene expression showed a positive correlation with OS, Here is the curve: